

Discussion 3

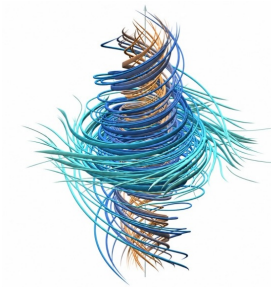


fig: A proposed vortex that
solves the Navier-Stokes
problem

This week's resources:

- 3Blue1Brown on Bayes Theorem YouTube video
- Steve Brunton: MLE with Examples YouTube video
- OpenAI found a solution to a Millenium Prize Problem with an AI model! link

1 Gaussian Mixture Clustering with Kangaroos and Berkeley Students (1D, Two Gaussians)

Consider a Gaussian Mixture Model (GMM) with two components in one dimension. Let $z_n \in \{K, B\}$ be the latent (unobserved) class of the n -th jumper (K = Kangaroo and B = Berkeley student) and $x_n \in \mathbb{R}$ be the observed jump height (in meters).

Assume that the classes occur with equal probability and that, conditional on the class, jump heights are Gaussian with class-dependent means μ_K and μ_B (the average jump heights of Kangaroos and Berkeley Students, respectively), and a shared variance σ^2 (the variability in jump heights).

(a) Write down the following quantities explicitly.

- (i) The **prior** probability that a jumper is a Kangaroo or Berkeley student before observing the jump height: $p(z_n)$.

Solution: Prior:

$$p(z_n = K) = p(z_n = B) = \frac{1}{2}$$

- (ii) The **likelihood** of the jump height conditioned on whether the jumper is a Kangaroo or Berkeley student: $p(x_n | z_n)$.

Solution: Conditional:

$$p(x_n | z_n = k) = \mathcal{N}(x_n; \mu_k, \sigma^2)$$

- (iii) The the **marginal** probability that a certain jump height is observed: $p(x_n)$.

Solution: Marginal:

$$p(x_n) = \sum_{k \in \{K, B\}} p(x_n | z_n = k) p(z_n = k) = \frac{1}{2} \mathcal{N}(x_n; \mu_K, \sigma^2) + \frac{1}{2} \mathcal{N}(x_n; \mu_B, \sigma^2)$$

- (iv) The **posterior** probability of a jumper being a Kangaroo or Berkeley student after observing the jump height: $p(z_n | x_n)$.

Solution: Posterior:

$$p(z_n = k | x_n) = \frac{p(x_n | z_n = k) p(z_n = k)}{p(x_n)}$$

(b) Recall the K-means objective:

$$J(\{r_{nk}\}, \{\mu_k\}) = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \|x_n - \mu_k\|^2$$

where μ_k is the mean of cluster k and $r_{nk} \in \{0, 1\}$ is a binary indicator variable that describes which of the K clusters the data point x_n is assigned to. The K-means algorithm alternates between updating the cluster assignments and updating the cluster centers.

We modify this algorithm to generate GMMs in the following way:

- (1) **Update assignments:** Replace the hard assignments r_{nk} with soft assignments γ_{nk} , representing the probability that a point x_n belongs to cluster k .
- (2) **Update parameters:** Rather than selecting cluster centers that minimize the squared distances of data points to their assigned cluster centers, select the parameters θ that maximize the Q function:

$$Q(\theta) = \sum_{n=1}^N \sum_{k=1}^K \gamma_{nk} \ln p(x_n, z_n = k \mid \theta).$$

- (i) Identify which probability from part (a) corresponds to γ_{nk} . Verify $\sum_k \gamma_{nk} = 1$.

Solution:

- (ii) The probability of assigning point x_n to class k is equivalent to the posterior probability $p(z_n = k \mid x_n)$. Since $p(z_n = k \mid x_n)$ is a valid probability distribution, summing over its domain $\{k\}_{k=1}^K$ will yield 1.
- (iii) Show that the Q function is actually the expected log-likelihood of our observed data, $X = (x_1, \dots, x_N)$, and the (unobserved) latent vector, $Z = (z_1, \dots, z_N)$ under the posterior probability of Z .¹

Solution: Using the probability we identified in part (i), the Q function is

$$Q(\theta) = \sum_{n=1}^N \sum_{k=1}^K p(z_n = k \mid x_n) \ln p(x_n, z_n = k \mid \theta) \quad (1)$$

$$= \sum_{n=1}^N \mathbb{E}_{z_n \sim p(z_n \mid x_n)} [\ln p(x_n, z_n \mid \theta)] \quad (2)$$

$$= \mathbb{E}_{Z \sim p(Z \mid X)} \left[\sum_{n=1}^N \ln p(x_n, z_n \mid \theta) \right] \quad (3)$$

$$= \mathbb{E}_{Z \sim p(Z \mid X)} \left[\ln \left(\prod_{n=1}^N p(x_n, z_n \mid \theta) \right) \right] \quad (4)$$

¹We optimize the joint likelihood because it is much more tractable to compute while still optimizing the marginal likelihood. See this page for more information.

$$= \mathbb{E}_{Z \sim p(Z|X)} \left[\ln p(X, Z | \theta) \right] \quad (5)$$

- (iv) Derive the maximum likelihood estimate for μ_k using the *expected* log-likelihood, $Q(\theta)$, treating γ_{nk} as constants.

Solution: To find the MLE estimate $\hat{\mu}_k$, we can take the derivative of the objective function with respect to μ_k and set it equal to 0. Before actually computing the derivative, it's helpful to expand out the probabilities and remove terms that don't directly depend on μ_k :

$$\frac{\partial Q(\theta)}{\partial \mu_k} = \frac{\partial}{\partial \mu_k} \sum_{n=1}^N \sum_{k'=1}^K \gamma_{nk'} \cdot \ln p(x_n, z_n = k' | \theta) \quad (6)$$

$$= \frac{\partial}{\partial \mu_k} \sum_{n=1}^N \gamma_{nk} \cdot \ln p(x_n, z_n = k | \theta) \quad (7)$$

$$= \frac{\partial}{\partial \mu_k} \sum_{n=1}^N \gamma_{nk} \cdot \ln (p(x_n | z_n = k, \theta) p(z_n = k)) \quad (8)$$

$$= \frac{\partial}{\partial \mu_k} \sum_{n=1}^N \gamma_{nk} \cdot (\ln p(x_n | z_n = k, \theta) + \ln p(z_n = k)) \quad (9)$$

$$= \frac{\partial}{\partial \mu_k} \sum_{n=1}^N \gamma_{nk} \cdot (\ln \mathcal{N}(x_n | \mu_k, \sigma^2)) \quad (10)$$

$$= \frac{\partial}{\partial \mu_k} \sum_{n=1}^N \gamma_{nk} \cdot -\frac{(x_n - \mu_k)^2}{2\sigma^2} \quad (11)$$

Taking the derivative yields

$$\frac{\partial Q(\theta)}{\partial \mu_k} = \sum_{n=1}^N \gamma_{nk} \frac{x_n - \mu_k}{\sigma^2} \quad (12)$$

$$(13)$$

Now we set this derivative equal to zero and solve for μ_k :

$$\begin{aligned} \sum_{n=1}^N \gamma_{nk} \mu_k &= \sum_{n=1}^N \gamma_{nk} x_n \\ \mu_k &= \frac{\sum_{n=1}^N \gamma_{nk} x_n}{\sum_{n=1}^N \gamma_{nk}} \end{aligned}$$

Notice this is very similar to how we estimated the cluster centers before, but now our center is a sum of sample points weighted by their posterior probabilities (γ_{nk}).

2 Binomial MLE with Misclassification

(a) You run m independent experiments. In each experiment, you perform n independent Bernoulli trials with unknown true success probability p . However, your measurement device is imperfect:

- A true success is recorded as a success with probability $1 - q$ (and as a failure with probability q).
- A true failure is recorded as a success with probability r (and as a failure with probability $1 - r$).

Let $X_i \sim \text{Bin}(n, p)$ denote the unobserved number of true successes in experiment i , and let Y_i be the observed number of recorded successes. Recall that the probability mass function for the binomial random variable $X_i \sim \text{Bin}(n, p)$ is

$$p(x_i) = \binom{n}{x_i} p^{x_i} (1 - p)^{n - x_i}.$$

Derive the likelihood for the observations $\mathcal{D} = \{y_i\}_{i=1}^m$. *Hint: Your answer should be a function of the true success probability p .*

Solution:

The variable Y_i is also a binomial random variable, where n is the number of trials, but the probability of an observed success is now a function of p :

$$\pi(p) = (1 - q)p + r(1 - p) = r + (1 - q - r)p.$$

Therefore, we can express the probability mass function parameterized by p :

$$p(y_i|p) = \binom{n}{y_i} \pi(p)^{y_i} (1 - \pi(p))^{n - y_i}.$$

Since experiments are independent, the likelihood of our full dataset

$$p(\mathcal{D}|p) = \prod_{i=1}^m \binom{n}{y_i} \pi(p)^{y_i} (1 - \pi(p))^{n - y_i}.$$

(b) Find the MLE $\hat{p}$ in closed form. Show all steps.

Solution: We want to find an expression for p that maximizes $p(\mathcal{D}|p)$. Let's begin with simplifying this likelihood by defining $s := \sum_{i=1}^m y_i$ and writing

$$p(\mathcal{D}|p) = \left(\prod_{i=1}^m \binom{n}{y_i} \right) \pi(p)^s (1 - \pi(p))^{mn - s}.$$

Notice that the first product does not depend on p , so we have

$$p(\mathcal{D}|p) \propto \pi(p)^s (1 - \pi(p))^{mn-s}.$$

Now, as we often do in MLE, we will look at the log likelihood:

$$\ln p(\mathcal{D}|p) \propto s \ln \pi(p) + (mn - s) \ln (1 - \pi(p)).$$

To find that expression for p that maximizes the log likelihood, notice that $\pi(p)$ is a linear function of p and the likelihood is a concave function of $\pi(p)$. Therefore, we can find the maximizer by taking the derivative with respect to p using the chain rule:

$$\frac{\partial}{\partial p} \ln p(\mathcal{D}|p) = \frac{\partial}{\partial \pi(p)} \ln p(\mathcal{D}|p) \frac{\partial}{\partial p} \pi(p).$$

Computing each factor in this product,

$$\frac{\partial}{\partial \pi(p)} \ln p(\mathcal{D}|p) = \frac{s}{\pi(p)} - \frac{mn - s}{1 - \pi(p)}, \quad \frac{\partial}{\partial p} \pi(p) = 1 - q - r.$$

Therefore,

$$\frac{\partial}{\partial p} \ln p(\mathcal{D}|p) = \left(\frac{s}{\pi(p)} - \frac{mn - s}{1 - \pi(p)} \right) (1 - q - r).$$

Now we need to set this expression equal to zero and solve for the critical points:

$$\begin{aligned} \frac{\partial}{\partial p} \ln p(\mathcal{D}|p) = 0 &\implies \frac{s}{\pi(p)} - \frac{mn - s}{1 - \pi(p)} = 0 \implies \\ s(1 - \pi(p)) &= (mn - s)\pi(p) \implies s = mn\pi(p) \implies \pi(p) = \frac{s}{mn} \end{aligned}$$

Now let's plug in our expression for $\pi(p)$ so we can solve for p :

$$r + (1 - q - r)p = \frac{s}{mn} \implies \boxed{p = \frac{\frac{s}{mn} - r}{1 - q - r}}.$$

3 Probability Review

An incoming email is spam with prior $p(S) = 0.2$ and not spam with $p(\bar{S}) = 0.8$. Two independent filters flag spam:

$$p(F_1=1 \mid S) = 0.9, \quad p(F_1=1 \mid \bar{S}) = 0.1, \quad p(F_2=1 \mid S) = 0.8, \quad p(F_2=1 \mid \bar{S}) = 0.05,$$

and F_1, F_2 are independent *given* the class (S or $\bar{S}$).

(a) What is the probability that both filters flag an email as spam?

Solution: We want to find $p(F_1 = 1, F_2 = 1)$. Conditional independence tells us

$$p(F_1 = 1, F_2 = 1 \mid c) = p(F_1 = 1 \mid c) p(F_2 = 1 \mid c).$$

We can then use the definition of conditional probability to find

$$p(F_1 = 1, F_2 = 1, c) = p(F_1 = 1, F_2 = 1 \mid c) p(c).$$

Finally, we can marginalize over c using the sum rule to get

$$p(F_1 = 1, F_2 = 1) = \sum_{c \in \{S, \bar{S}\}} p(F_1 = 1, F_2 = 1, c).$$

Putting this all together gives us

$$p(F_1=1, F_2=1) = \sum_{c \in \{S, \bar{S}\}} p(c) p(F_1=1 \mid c) p(F_2=1 \mid c) \quad (14)$$

$$= 0.2 \cdot (0.9 \cdot 0.8) + 0.8 \cdot (0.1 \cdot 0.05) \quad (15)$$

$$= 0.144 + 0.004 \quad (16)$$

$$= \boxed{0.148} \quad (17)$$

(b) Given that both filters flag an email, what is the probability of the email being spam? (You can leave your answer as an unsimplified fraction.)

Solution:

We wish to calculate $p(S \mid F_1 = 1, F_2 = 1)$. We can utilize Bayes' Theorem, which gives us

$$p(S \mid F_1 = 1, F_2 = 1) = \frac{p(F_1 = 1, F_2 = 1 \mid S) p(S)}{p(F_1 = 1, F_2 = 1)}.$$

The denominator comes from part (a), $p(S)$ is given to us as 0.2, and $p(F_1 = 1, F_2 = 1 \mid S) = p(F_1 = 1 \mid S) p(F_2 = 1 \mid S) = 0.9(0.8) = 0.72$ by conditional independence. Therefore, we have

$$p(S \mid F_1 = 1, F_2 = 1) = \frac{0.72(0.2)}{0.148} = \boxed{\frac{0.144}{0.148} \approx 0.973.}$$